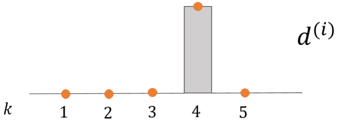
Introduction to Machine Learning

Information Theory for Machine Learning



Learning goals

- Minimizing KL = maximizing log-likelihood
- Minimizing KL = minimizing cross-entropy
- Minimizing CE between modeled and observed probabilities = log-loss minimization

KL VS MAXIMUM LIKELIHOOD

Minimizing KL between the true distribution $p(x)$ and approximating model $q(x|\boldsymbol{\theta})$ is equivalent to maximizing the log-likelihood.

$$
\begin{aligned}
D_{KL}(p\|q_{\boldsymbol{\theta}}) &= \mathbb{E}_{X\sim p}\left[\log \frac{p(x)}{q(x|\boldsymbol{\theta})}\right] \\
&= \mathbb{E}_{X\sim p}\log p(x) - \mathbb{E}_{X\sim p}\log q(x|\boldsymbol{\theta})
\end{aligned}
$$

as first term above does not depend on $\boldsymbol{\theta}$. Therefore,

$$
\begin{aligned}
\arg\min_{\boldsymbol{\theta}} D_{KL}(p\|q_{\boldsymbol{\theta}}) &= \arg\min_{\boldsymbol{\theta}} -\mathbb{E}_{X\sim p}\log q(x|\boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{X\sim p}\log q(x|\boldsymbol{\theta})
\end{aligned}
$$

For a finite dataset of n samples from p, this is approximated as

$$
\arg\max_{\boldsymbol{\theta}} \mathbb{E}_{X\sim p}\log q(x|\boldsymbol{\theta}) \approx \arg\max_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n}\log q(\mathbf{x}^{(i)}|\boldsymbol{\theta}) \,.
$$

This also directly implies an equivalence to risk minimization!

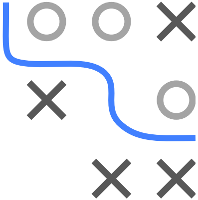
KL VS CROSS-ENTROPY

From this here we can see much more:

$$\arg\min_\theta D_{KL}(p \| q_\theta) = \arg\min_\theta -\mathbb{E}_{X\sim p} \log q(x|\theta) = \arg\min_\theta H(p \| q_\theta)$$

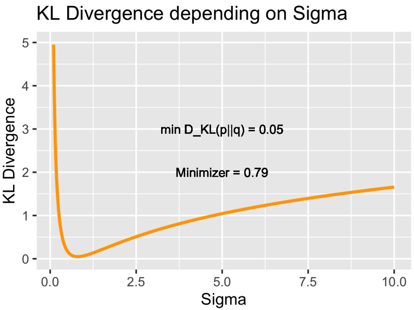
- So minimizing KL is the same as minimizing CE, is the same as maximum likelihood!
- We could now motivate CE as the "relevant" term that you have to minimize when you minimize KL - after you drop $\mathbb{E}_p \log p(x)$, which is simply the neg. entropy H(p)!
- Or we could say: CE between p and q is simply the expected negative log-likelihood of q, when our data comes from p!

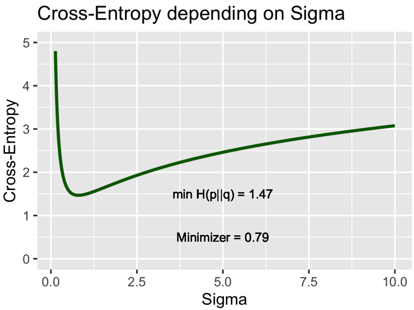
KL VS CROSS-ENTROPY EXAMPLE



Let $p(x) = N(0,1)$ and $q(x) = LP(0,\sigma)$ and consider again

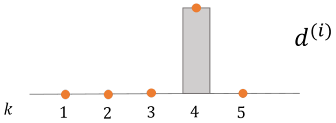
$$\arg\min_{\boldsymbol{\theta}} D_{KL}(p\|q_{\boldsymbol{\theta}}) = \arg\min_{\boldsymbol{\theta}} -\mathbb{E}_{X\sim p} \log q(x|\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} H(p\|q_{\boldsymbol{\theta}})$$

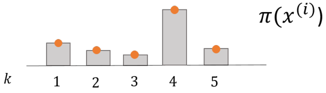




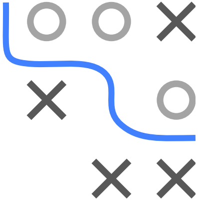
CROSS-ENTROPY VS. LOG-LOSS

- Consider a multi-class classification task with dataset $\mathcal{D} = \left(\left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(n)}, y^{(n)}\right)\right)$.
- For g classes, each $y^{(i)}$ can be one-hot-encoded as a vector $d^{(i)}$ of length g. $d^{(i)}$ can be interpreted as a categorical distribution which puts all its probability mass on the true label $y^{(i)}$ of $\mathbf{x}^{(i)}$.
- $\pi(\mathbf{x}^{(i)}|\theta)$ is the probability output vector of the model, and also a categorical distribution over the classes.





CROSS-ENTROPY VS. LOG-LOSS / 2



To train the model, we minimize KL between $d^{(i)}$ and $\pi(\mathbf{x}^{(i)}|\boldsymbol{\theta})$:

$$\arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} D_{KL}(d^{(i)} \| \pi(\mathbf{x}^{(i)}|\boldsymbol{\theta})) = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} H(d^{(i)} \| \pi(\mathbf{x}^{(i)}|\boldsymbol{\theta}))$$

We see that this is equivalent to log-loss risk minimization!

$$\begin{aligned} R &= \sum_{i=1}^{n} H(d^{(i)} \| \pi_k(\mathbf{x}^{(i)}|\boldsymbol{\theta})) \\ &= \sum_{i=1}^{n} \left(- \sum_{k} d_k^{(i)} \log \pi_k(\mathbf{x}^{(i)}|\boldsymbol{\theta}) \right) \\ &= \sum_{i=1}^{n} \underbrace{\left(- \sum_{k=1}^{g} [y^{(i)} = k] \log \pi_k(\mathbf{x}^{(i)}|\boldsymbol{\theta}) \right)}_{\text{log loss}} \\ &= \sum_{i=1}^{n} (- \log \pi_{y^{(i)}}(\mathbf{x}^{(i)}|\boldsymbol{\theta})) \end{aligned}$$

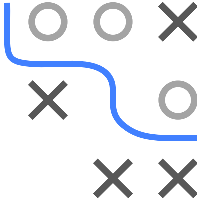
CROSS-ENTROPY VS. BERNOULLI LOSS

For completeness sake:
Let us use the Bernoulli loss for binary classification:

$$L(y, \pi(\mathbf{x})) = -y \log(\pi(\mathbf{x})) - (1 - y) \log(1 - \pi(\mathbf{x}))$$

If p represents a $\text{Ber}(y)$ distribution (so deterministic, where the true label receives probability mass 1) and we also interpret $\pi(\mathbf{x})$ as a Bernoulli distribution $\text{Ber}(\pi(\mathbf{x}))$, the Bernoulli loss $L(y, \pi(\mathbf{x}))$ is the cross-entropy $H(p \| \pi(\mathbf{x}))$.

ENTROPY AS PREDICTION LOSS



Assume log-loss for a situation where you only model with a constant probability vector π. We know the optimal model under that loss:

$$\pi_k = \frac{n_k}{n} = \frac{\sum\limits_{i=1}^{n} [y^{(i)} = 1]}{n}$$

What is the (average) risk of that minimal constant model?

$$\mathcal{R} = \frac{1}{n} \sum_{i=1}^{n} \left(- \sum_{k=1}^{g} [y^{(i)} = k] \log \pi_k \right) = - \frac{1}{n} \sum_{k=1}^{g} \sum_{i=1}^{n} [y^{(i)} = k] \log \pi_k$$

$$= - \sum_{k=1}^{g} \frac{n_k}{n} \log \pi_k = - \sum_{k=1}^{g} \pi_k \log \pi_k = H(\pi)$$

So entropy is the (average) risk of the optimal "observed class frequency" model under log-loss!